RNA sequencing

Paul Bertone

EMBL-EBI

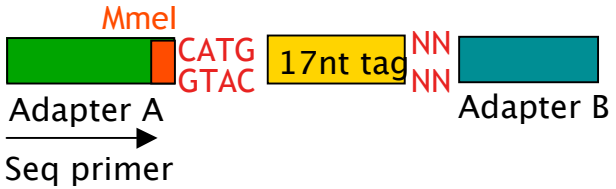
EBI is an Outstation of the European Molecular Biology Laboratory.

Solexa transcriptome sequencing

- Solexa data analysis and associated software development
  - Unbiased expression profiling
  - Tandem identification of expressed non-coding RNAs
  - MicroRNA identification and expression analysis
- Advantages over microarrays
  - Gene expression arrays don't capture unannotated transcripts
  - Tiling arrays are still expensive for large genomes (e.g. mammals)
  - Small RNAs are too short for stable hybridization
  - No fluorescence correction to account for, essentially zero background
- Current disadvantages
  - More expensive than standard expression arrays
  - More time consuming than any microarray technology
  - Some data analysis issues
    - No strand orientation information – sequencing a double-stranded product
    - Computing accurate transcript models, mapping reads to splice junctions
    - Contribution of high-abundance RNAs (eg ribosomal) could dilute the remaining transcript population; sequencing depth is important

Transcriptome sequencing methods

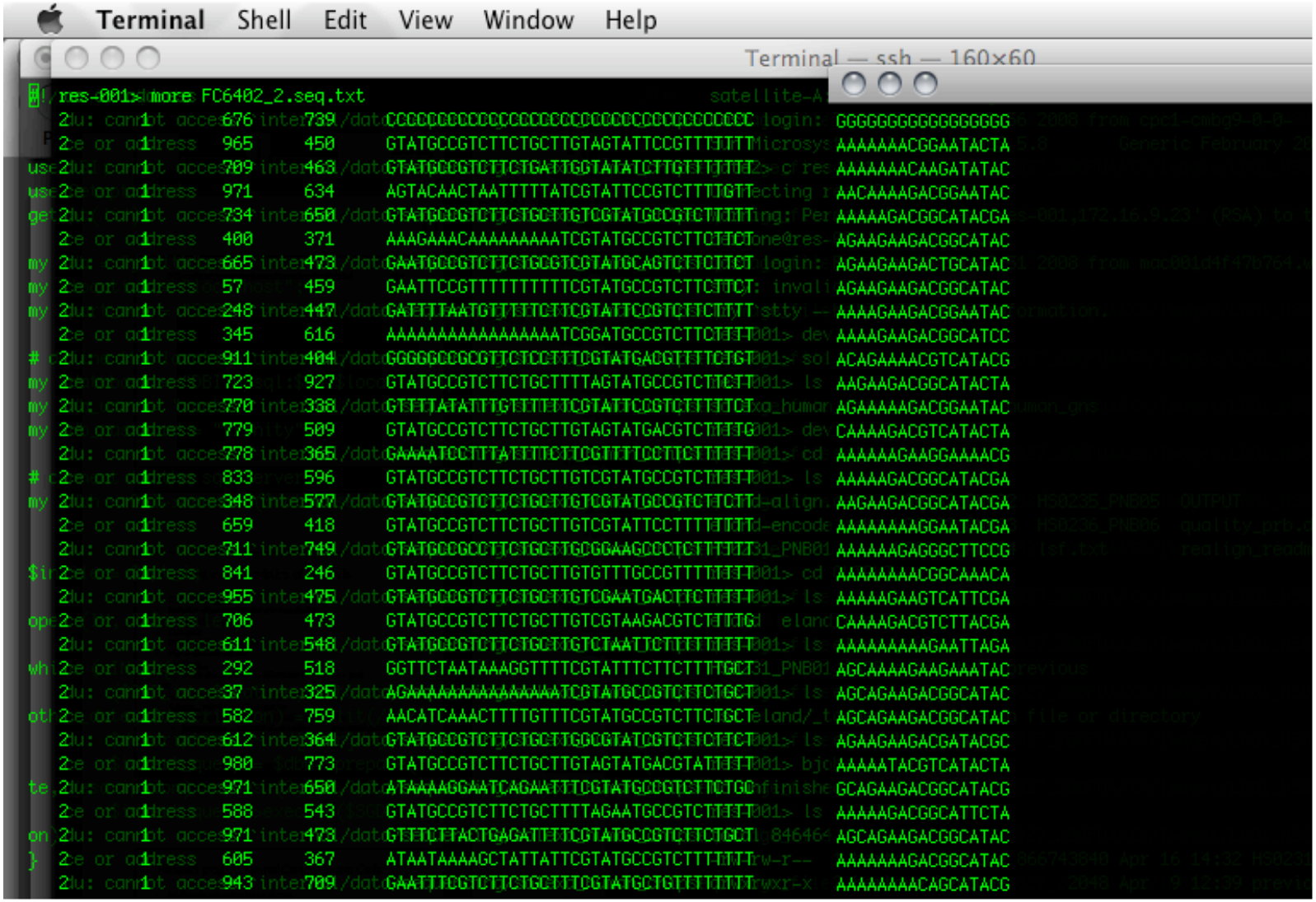
Method 1: variant of the LongSAGE protocol



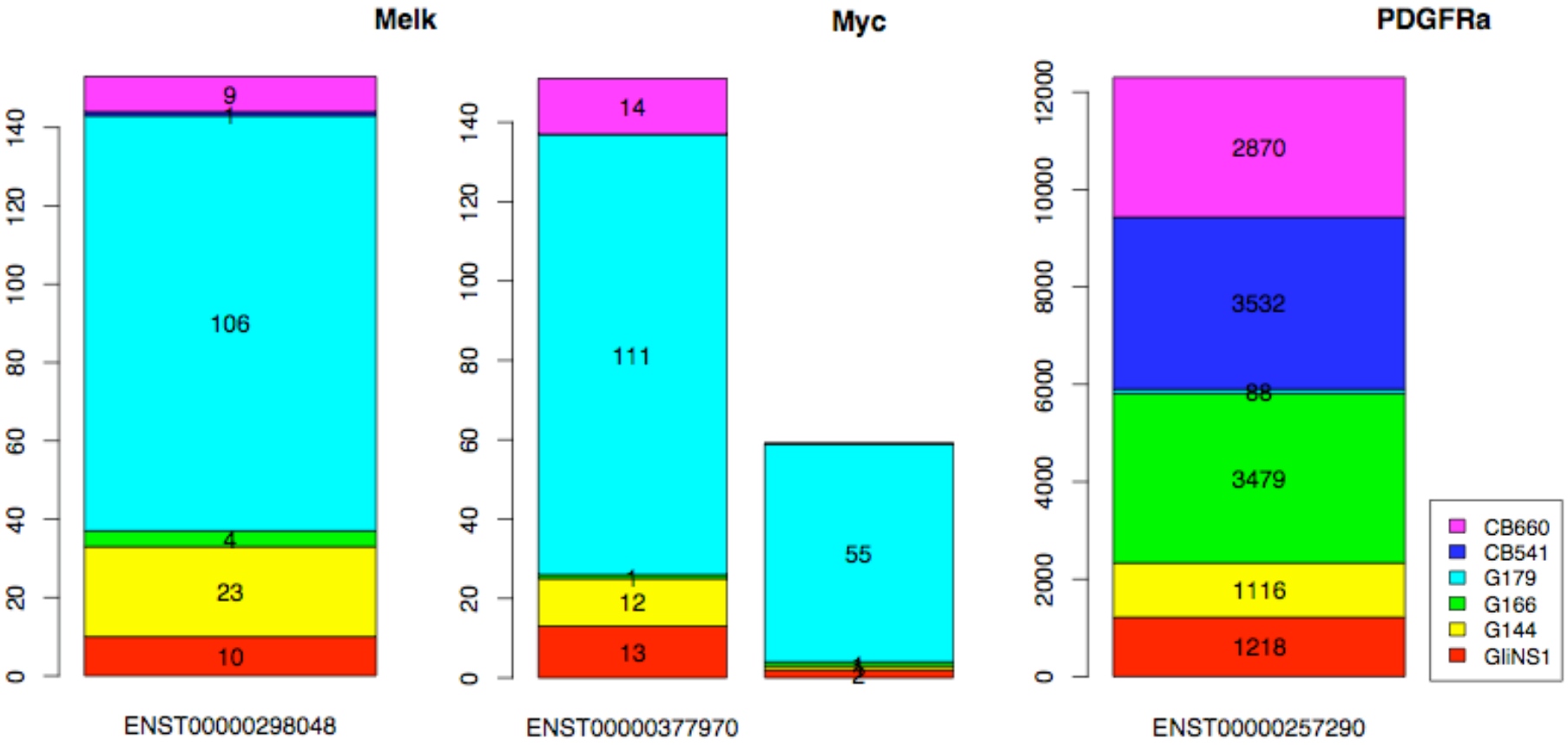
Poly-A RNA selection
Double strand cDNA synthesis on beads
NlaIII digestion to remove 5' portion of cDNAs
Ligation to 5' adapters containing a MmeI recognition site
MmeI digestion to remove the 3' portion of cDNA
This generates a 17nt tag (not including CATG)
Tags are ligated to a 3' adapter
The construct is PCR-amplified using primers homologous to 5' and 3' adapters
PCR products are purified and quantitated (e.g. with Agilent Bioanalyzer)
Load tag-adapter hybrids into flow cell lanes and sequence

- No concatenation of SAGE tags
- One tag is amplified and sequenced per flow cell cluster
- Read (tag) alignment is performed against a library of virtual tags

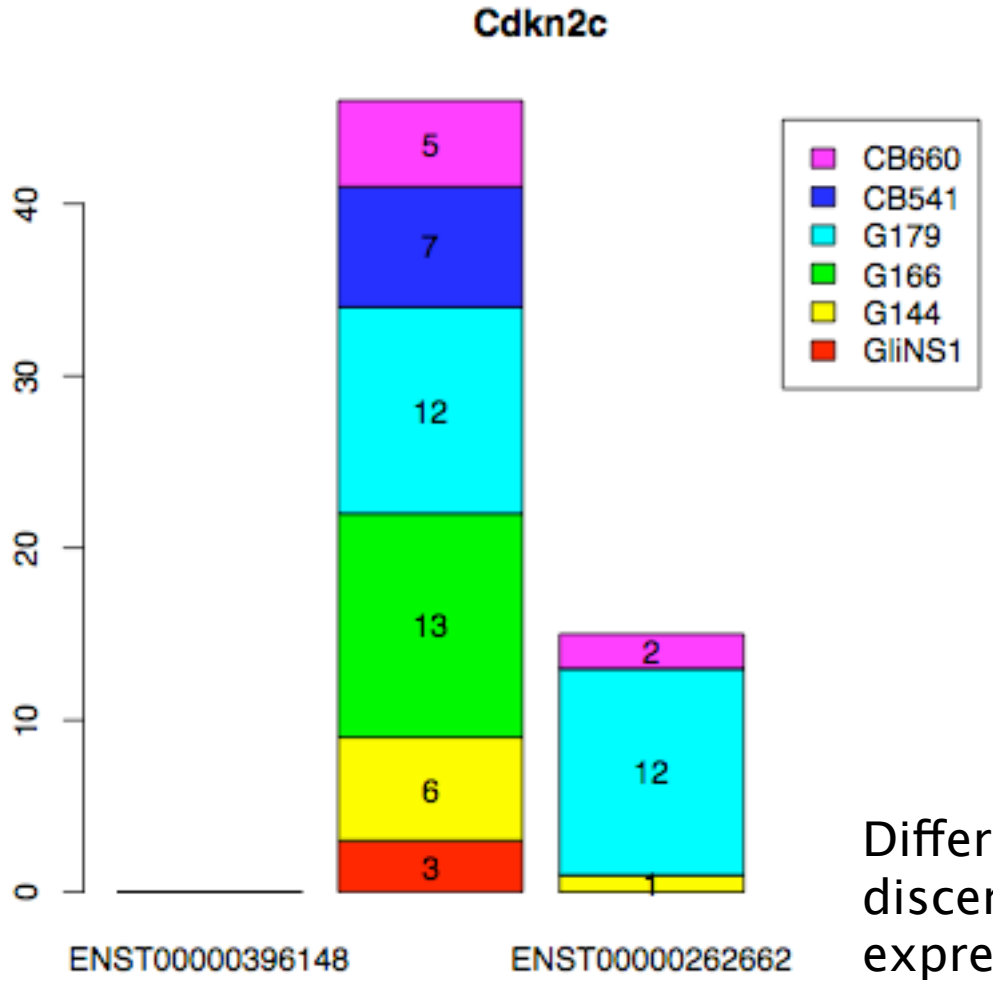
SAGE sequencing output



Solexa transcriptome sequencing



Solexa transcriptome sequencing



Differential read counts allow us to discern which transcript isoforms are expressed

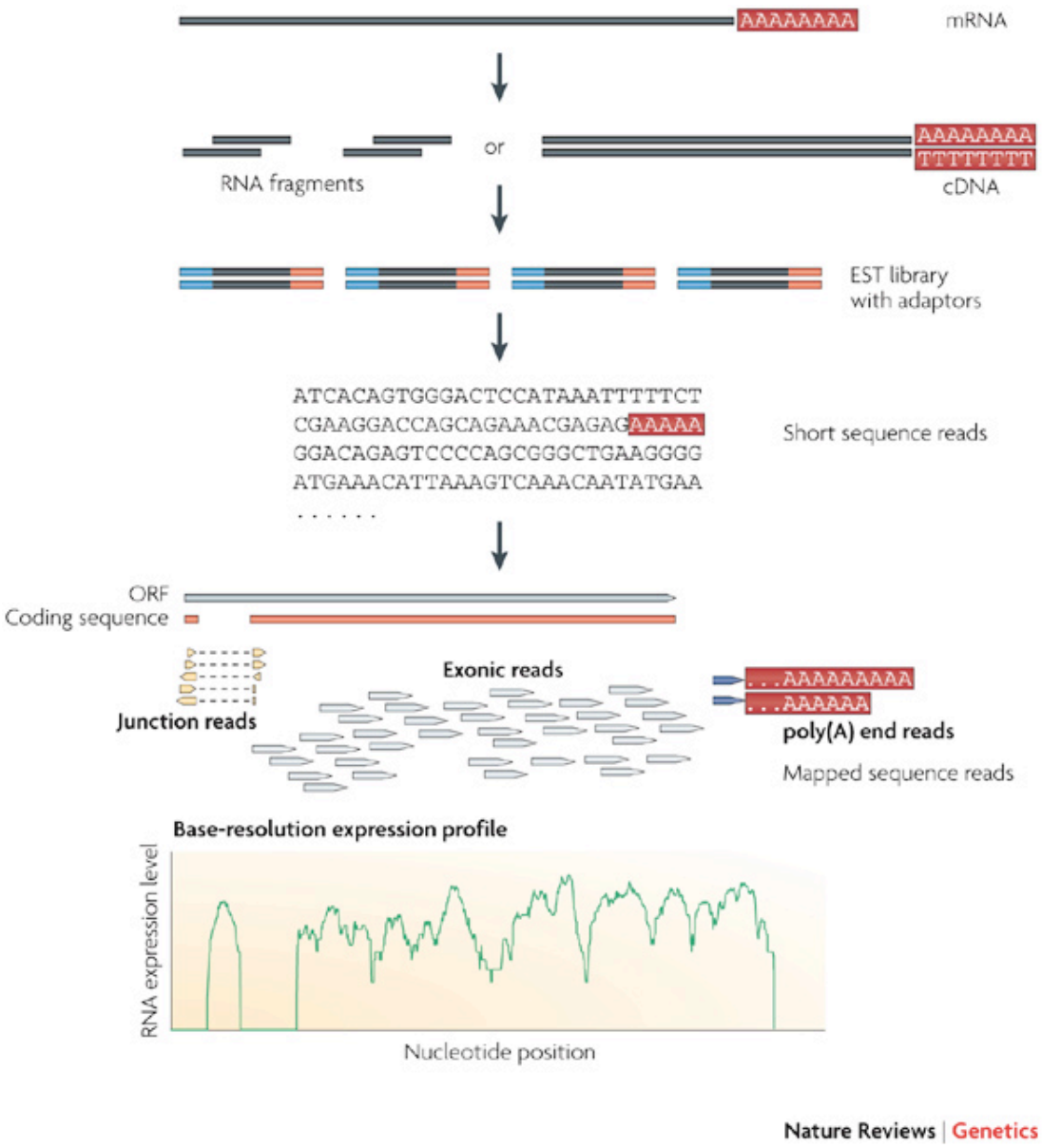
Features of SAGE analysis

- Complicated library construction
- Good at gene expression analysis
- Short reads (17nt), therefore low rate of unique alignments to reference genome
  - Reads are mapped to virtual tags instead
- Mostly limited to annotated genes
- Can get some information on novel transcripts (limited)

mRNA sequencing

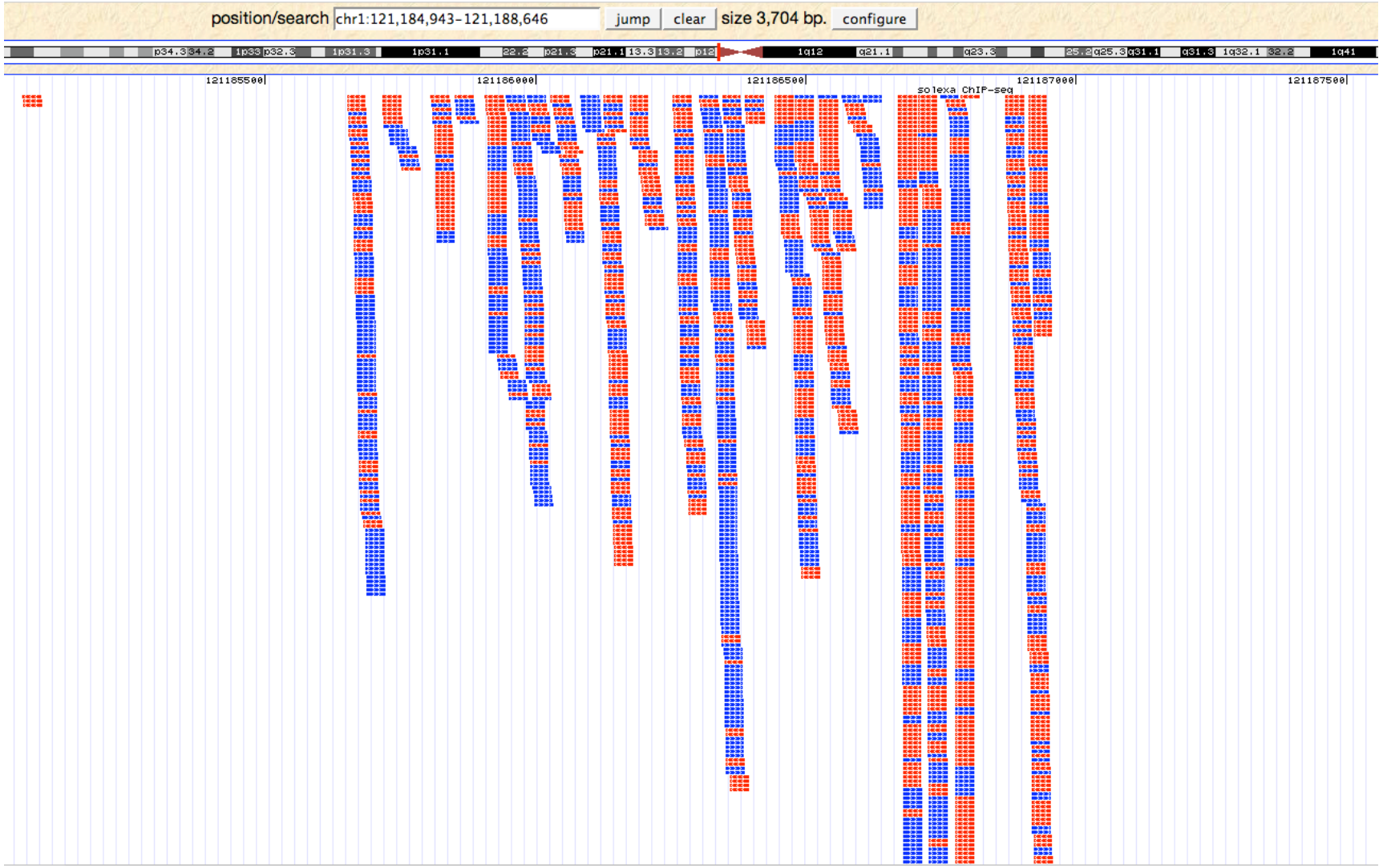
- Similar to SAGE analysis in terms of gene expression
- Simpler library construction
- Not limited to 17nt reads
  - Utilize full read length for alignment
  - Much better genome mapping
- Results are analogous to tiling array profiling
  - Reads map to individual transcript components
  - Ascertain splice variation as well as gene expression
  - Refine existing annotation of exons and UTRs
  - Identify non-coding RNAs

mRNA-seq protocol

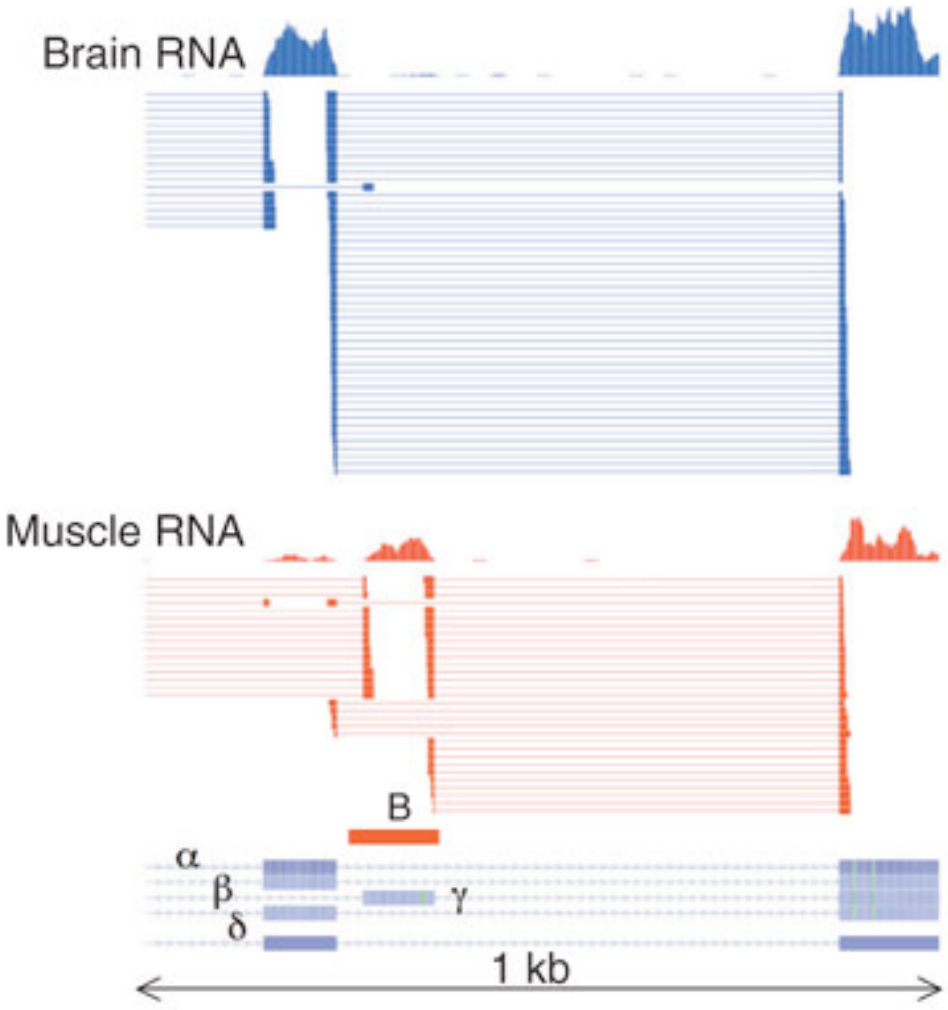


Nature Reviews | Genetics

Reads mapped to the human genome

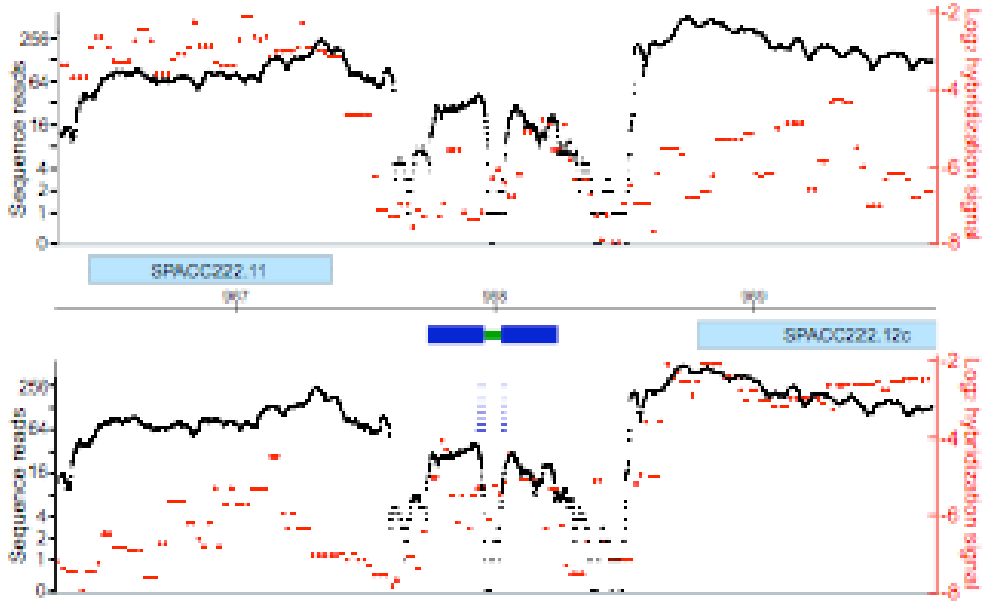


Alignment to exon splice junctions



Alignment reference should consider mature transcripts or exon junctions

Sequencing vs. tiling array hybridization



Red: tiling array hybridization signal (log2)

Black: sequencing reads (log)

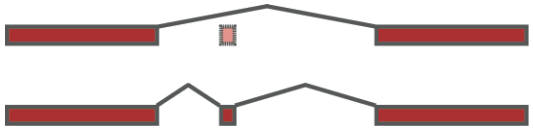
- Example comparison between Solexa WTSS and tiling array hybridization data (S. pombe, Bahler lab Sanger)

- Top image = sense strand; bottom image = antisense strand
- Light blue = annotated genes; Dark blue = new non-coding transcript; Green = intron

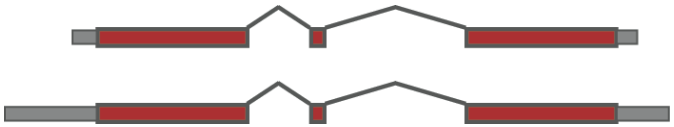
Novel Transcribed Regions: Possibilities

- Many areas of active transcription are observed outside annotated genes
  - Rare or low-abundance protein-coding transcripts

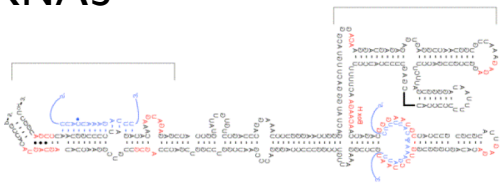
  - Unannotated exons from alternate splice products



  - Previously under-represented 3' and 5' UTRs

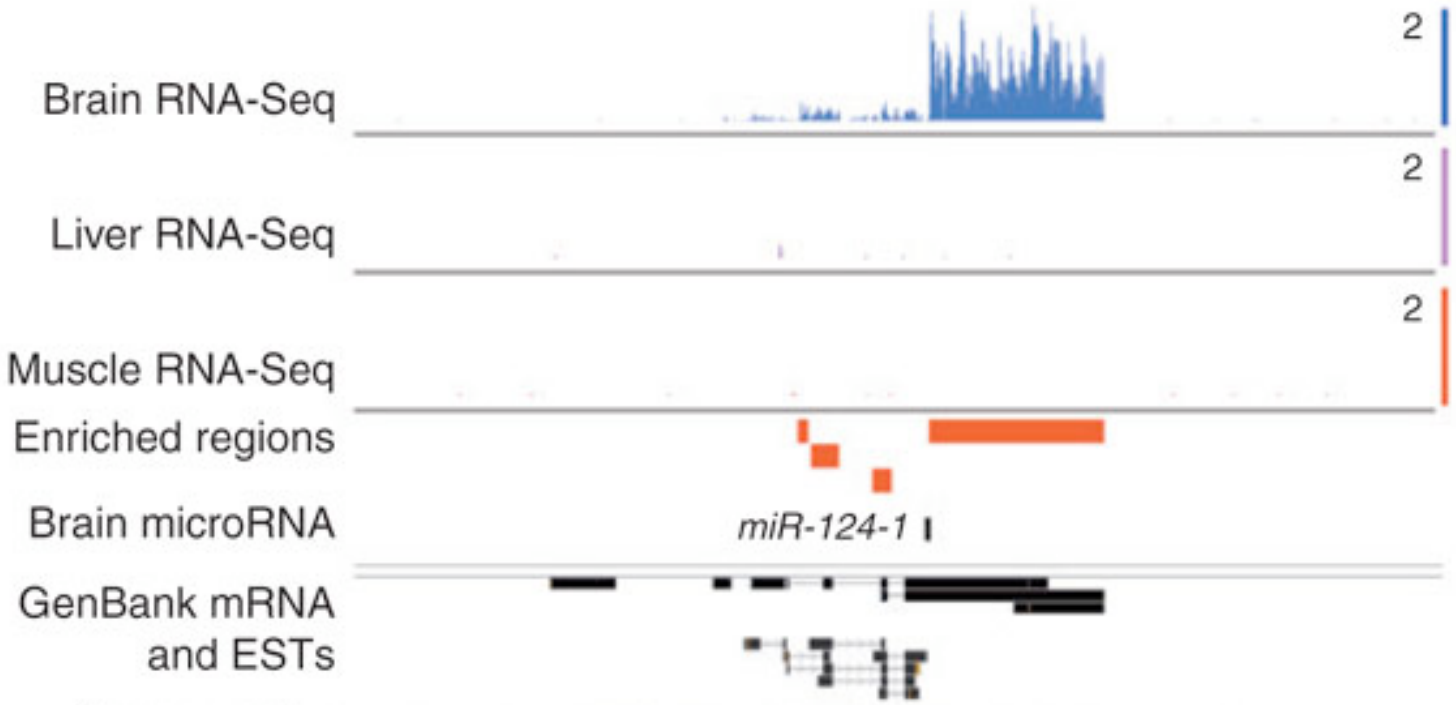


  - Noncoding RNAs



Splice variation, refinement of existing exon annotation

Detection of microRNA precursors



Protocol variations

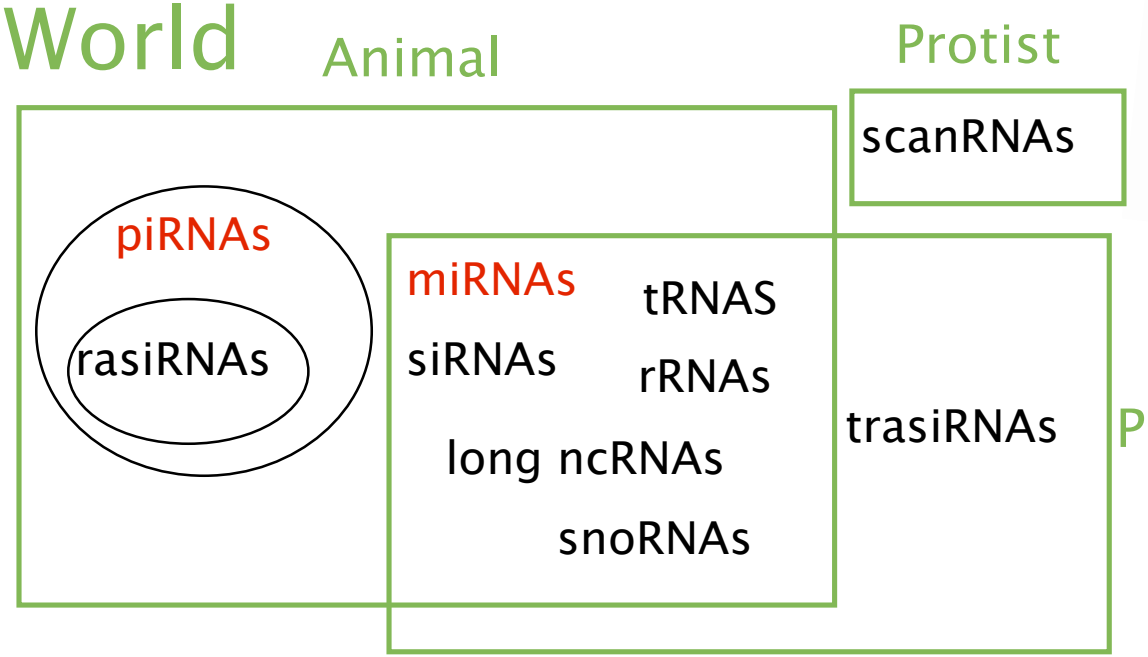
- Fragmentation methods
  - RNA: nebulization, hydrolysis
  - cDNA: sonication, Dnase I treatment
- Depletion of highly abundant transcripts
  - e.g. RiboMinus – others?
- Oligo-dT selection for poly(A)+ transcripts vs total RNA
- Coverage issues
  - What is the sequencing depth required?
- Strand specificity
  - Most RNA sequencing is not strand-specific
  - Currently working with Vladimir Benes and Lars Steinmetz on new protocols for this

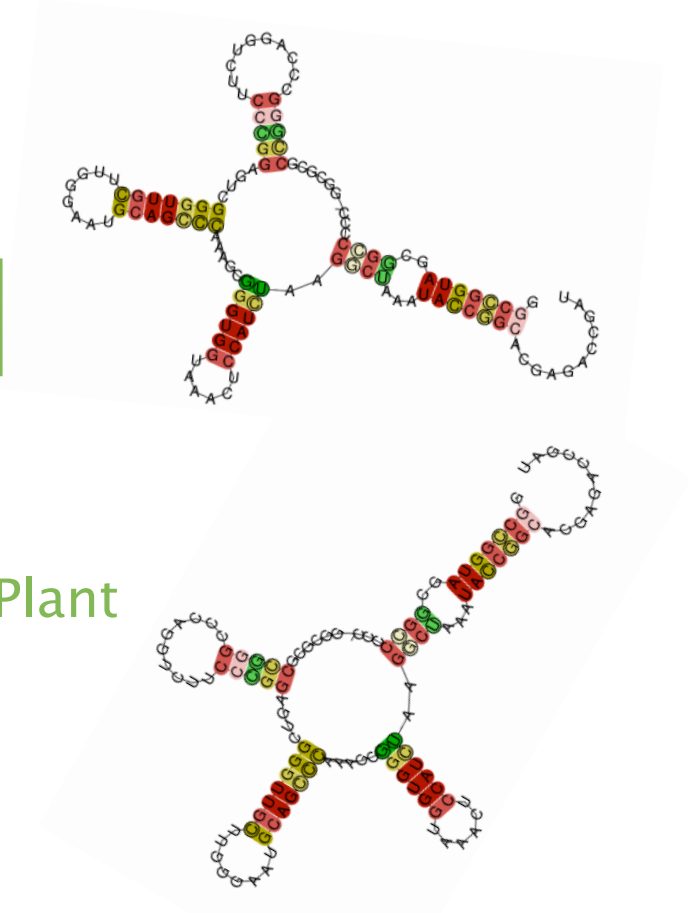
Specialized RNA-seq applications

- Small RNA sequencing
  - microRNAs
  - piRNAs
  - endo-siRNAs

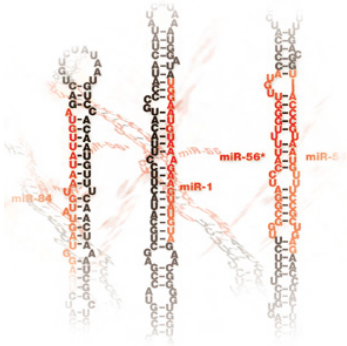
- Identification of RNAs associated with protein complexes (e.g. Ago2)
  - Immunoprecipitation of RNA-bound protein complexes
  - Proteinase K digestion, purification of nucleic acids for sequencing

The Non-coding RNA World

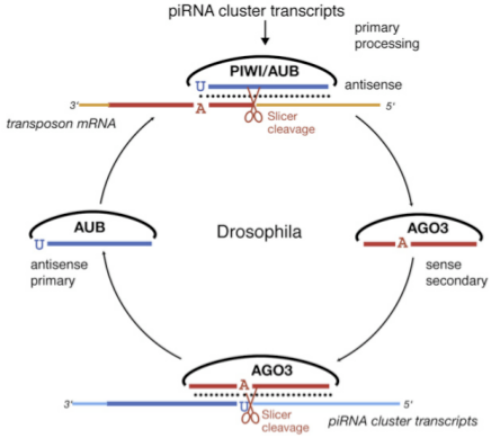




- Growing number of non-coding RNA classes categorized by many different features (e.g. function, length, secondary structures, expression tissues, species, etc.)
- For my projects I am focusing on short regulatory non-coding RNAs..
- ..paying particular attention to the microRNA and piwiRNA classes



microRNAs and piwiRNAs



Differences

- miRNAs are generally shorter (~21-23nt) than piRNAs (~24-30nt)
- miRNAs are Dicer-dependent
- miRNAs are processed from a dsRNA precursor with a known secondary structure (piRNAs?)
- Expression of piRNAs is thought to be restricted to the germline
- miRNAs bind to Argonaute clade whilst piRNAs to the Piwi clade of the Argonaute protein family

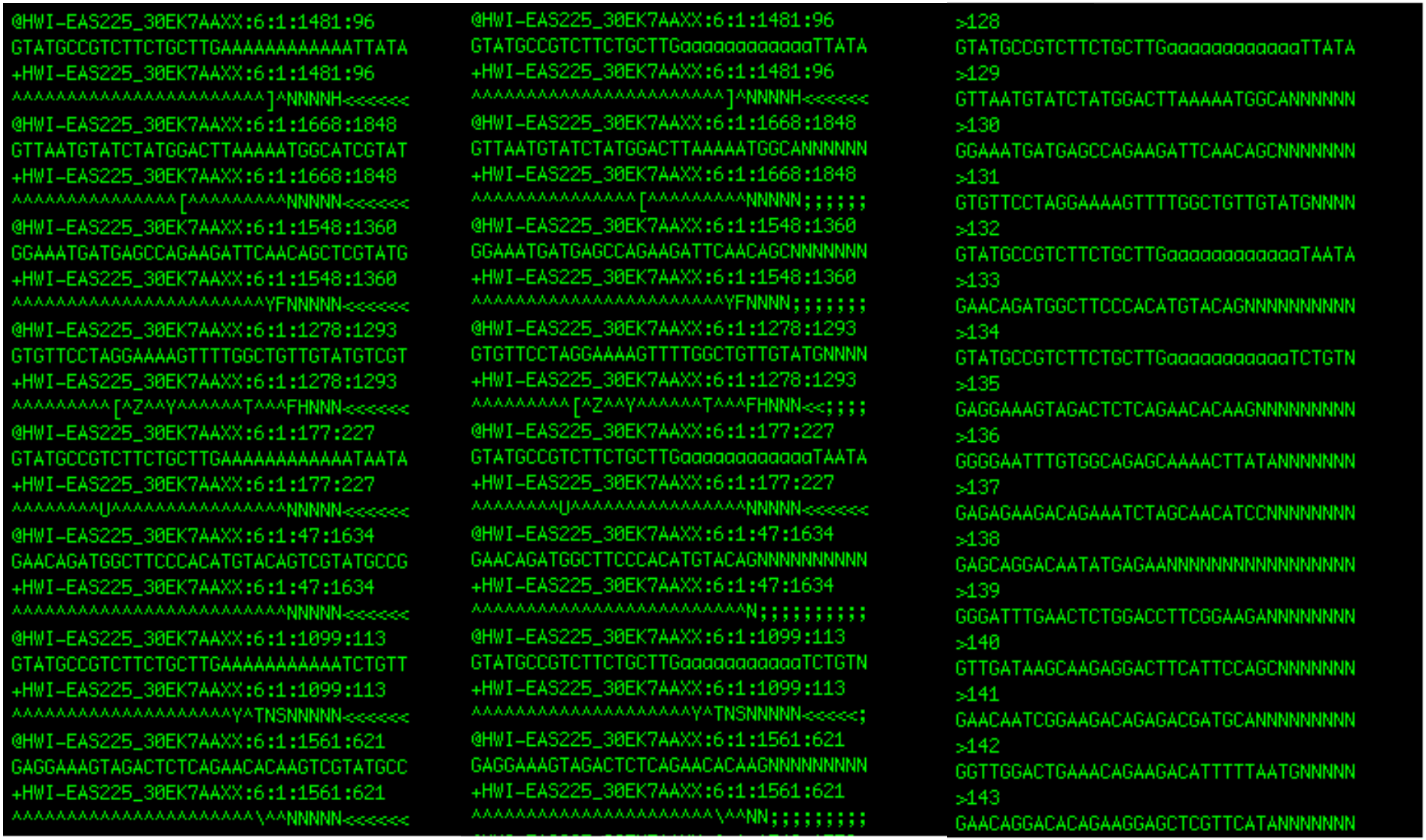
Similarities

- Both show a 5'Up preference
- Both show a 2'O-methyl modification at their 3' end (plant microRNAs only)

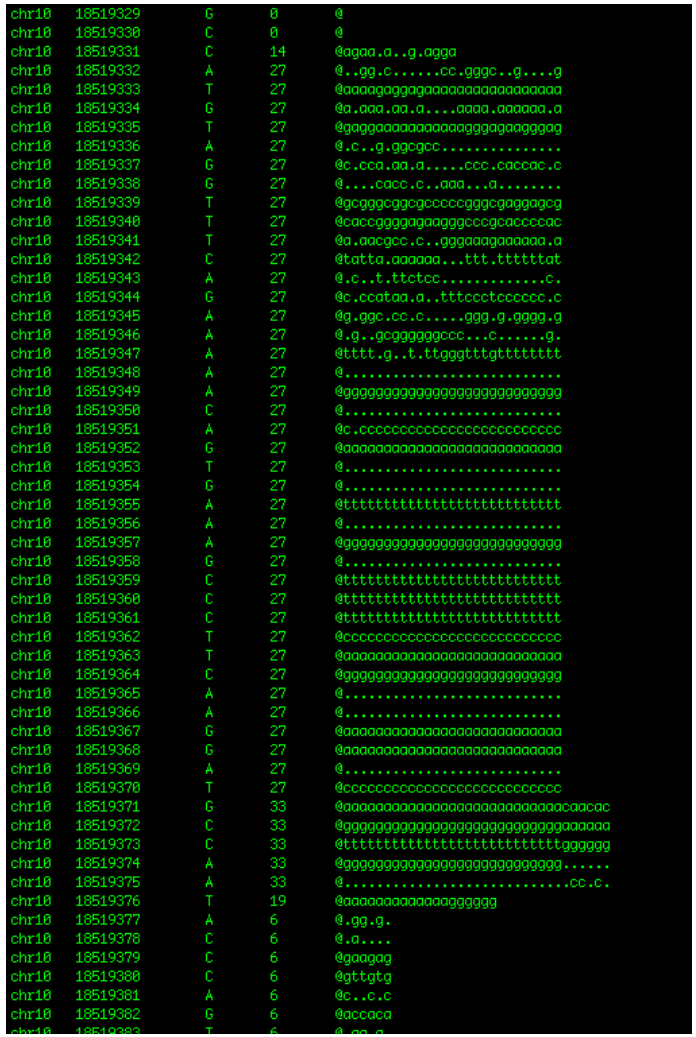
Differences in small RNA sequencing

- Size exclusion of total RNA
  - Selected to target particular species
  - e.g. 17–23nt for microRNAs, 25–32nt for piRNAs
  - 17–32nt can encompass both populations
- Direct ligation of adapters to RNA molecules
- Transcripts are typically shorter than the reads
  - Sequence into the adapters
  - Reveals strand specificity

Adapter masking, low-complexity filtering

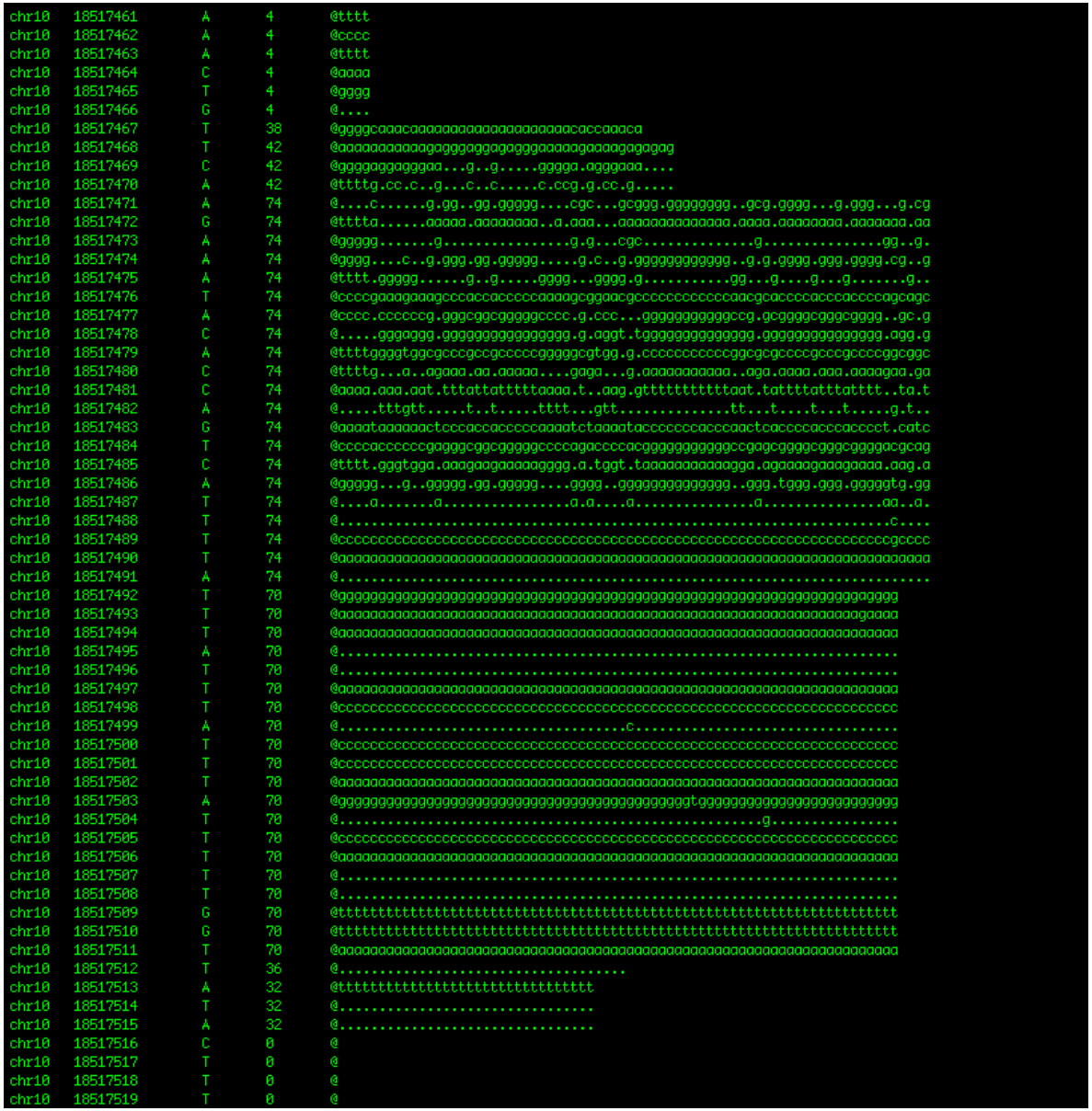


Aligned RNA reads from RNA-seq



Sample	KS35	KS45
Reads	**3,559,384**	**5,861,316**
Eland placement (total)	**2,429,078** (68%)	**4,109,776** (70%)
Unique, no mismatch	1,806,384	3,445,856
Unique, 1 mismatch	418,151	440,111
Unique, 2 mismatches	204,543	223,809

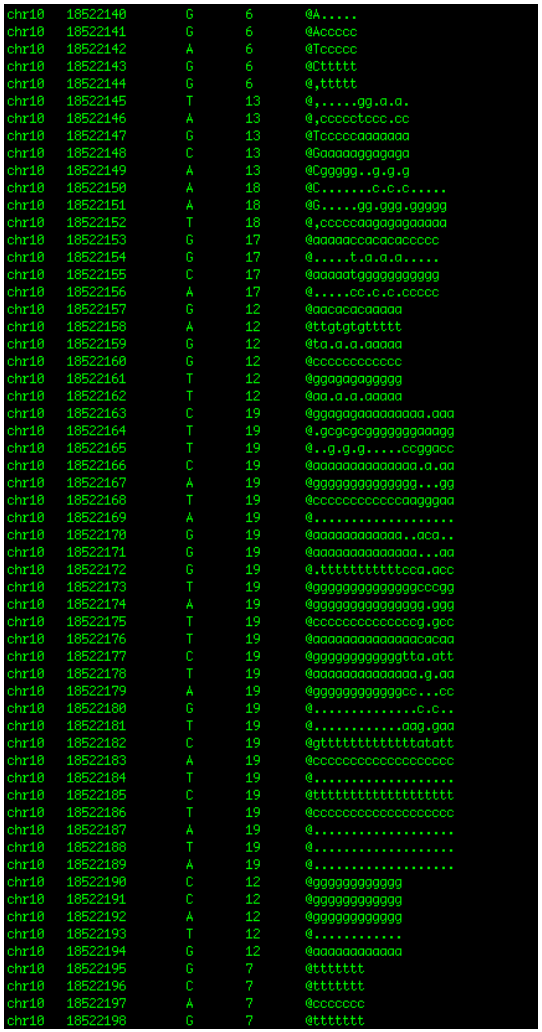
Read depth varies across different loci



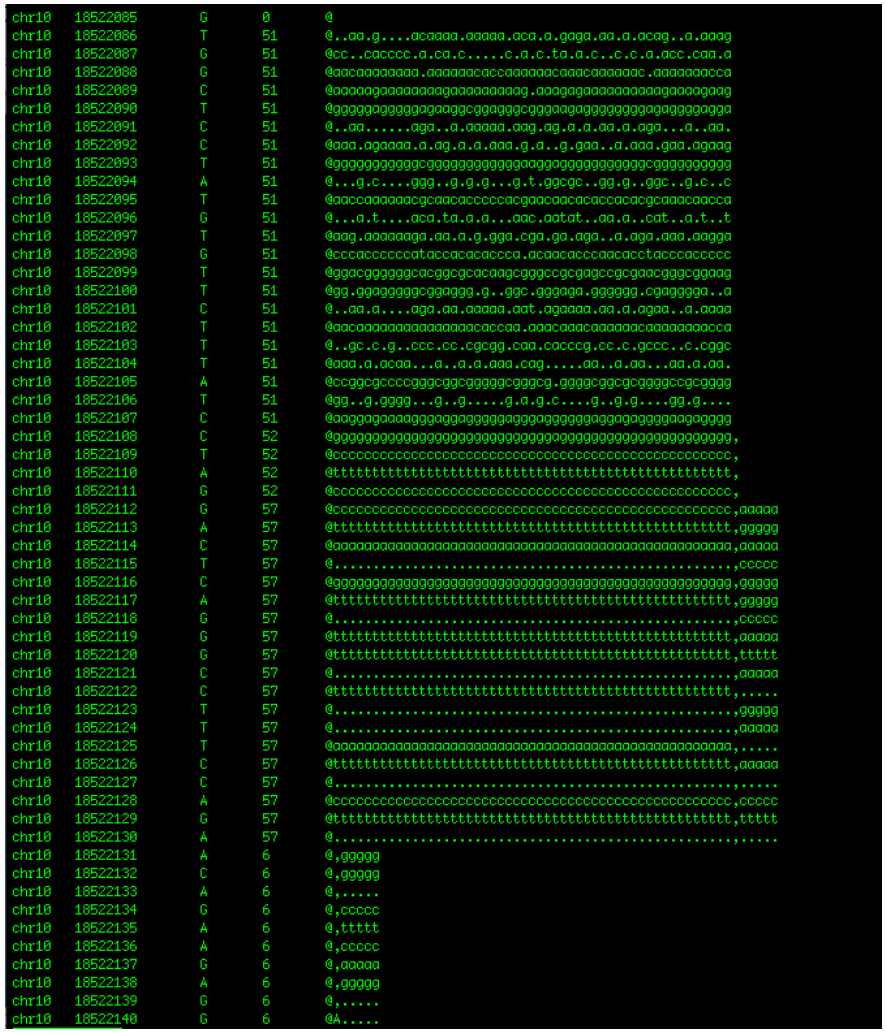
Transcriptional units from RNA-seq

```
>MM9:10:18509523-18509567
AGTAGGAAGTTATGGTATCTTTGGAAAGTCAGTTGTGTTAGCTGG
>MM9:10:18516039-18516083
TGGTGGAGTCTTCTTTTGGTTGTCATTGGGGACCTATAGAGGGCA
>MM9:10:18516279-18516359
ATGTGCTTAATTTTTGTATTCTGTATACTTTCCTGGCATTGTCTGATCACCTACATTCATTCATTCATTCATTCACC
>MM9:10:18516802-18516857
ACTTTCCTAGTCTACCTATGTCCATCTAACCTGCCTTCCTTCCCCACTTCAGAGAT
>MM9:10:18516888-18516932
ATGACCGCTCTAGCTGCTTCATGCTGTGAGGGGGCGGAGTCCAGG
>MM9:10:18516988-18517060
CTTCTTTAGACAGGCTGTACGTATCTGAGGCTCCACGTAAAGGGCCATCTGTAGAAGTGGAAGGGCTGAAGGA
>MM9:10:18517358-18517407
TATCTAGGCAGCATTTTTCTTATCAAGAAGAAAATCACTAATGAGAAATG
>MM9:10:18517447-18517515
GAGTATGGTCCCTTAAACTGTTCAAGAAATACACCAGTCATTTTATTTATTTATTTATTTTTGGTTATA
>MM9:10:18517618-18517724
TGAGCGTCTAACCTTCCTTTTCCAAGTATGCTCATTCTACGGGAGGAGAATTTATGCTTAGCTATGGCTTTACTTGGAAACATGGAAGCAAGATCCATGGAGGGTCC
>MM9:10:18517761-18517826
TTGAGTATACTGGTTGGCGTCTCCATGACCTCCCTGAACAACAGAAAAGGTGATTCCTCACAGTTA
>MM9:10:18517836-18517888
TTGTGTGAGATGCTTGTCCGACCTACTTGATCTTGGGGGCCAAGGAGGAATAT
>MM9:10:18517895-18517964
CTTTTTAATCCCTCTGACCACATTAGTATGGTCTCCAAGTATGGTTATTGAACACCCAGGATGCCACTGA
>MM9:10:18518002-18518149
TTTTTTTCCTCAAACAATCTTTGTAATTGTTAGGGAGAACAGGCCATTTATTTAAGCAGAGCTTTGACATCAATGTATTGAATTCTGACTGTTTACATACCTTATATAACTTGCTCAGACCCGTATAATTTGCTTAAGCTTTCCAACT
>MM9:10:18518187-18518268
GAGACATAATAACATTAACAAGCCAATAATCAACAAGACTAGCAGTGATGATCCTTTGAAATTTGGCATAGCAATCCTTGGC
>MM9:10:18518350-18518399
TAAATTACATGTATGCATACTGTTAAATTATGGGGCCTGTTAAGTTTCCA
>MM9:10:18518496-18518549
AGAACAAAAGTGGTCATAGTCCCAAAGATCATCAGGGGAGCAGGATCTATGGCA
>MM9:10:18518610-18518711
ATCTGTTGGTCTGGGAGCTTTCCTAAAGTCCCCAGGAGGACTTATGCACTACAGTAATCTTCAAGCATCTATCAAGCTTCAAGAACCAGCTCAATTAACTCT
>MM9:10:18518781-18518865
ATTTTAAAAAGTACATATATATATACACATACACACTAGGTGCTTCTTTTTAAAAGATCCTTGGGCCAAGAGATCACAATCTATG
>MM9:10:18518930-18519120
ACACCTTCTAGATCTTGGCTTCCTCTCCAAAGTGCCCTCTACAGAGTTAAATAGACATGGGTATGTGTACTTGTAGCCTCTTCTGACTCACTAGCAACCATGGAAGAAGGGGCAGCTTGTTCATAGCATTTGCTCTAAGGCGAATGCCCAAAGTCTCCCG
TAAGGCCATCAGAGAGCTCTTTACACACAAC
>MM9:10:18519161-18519278
TGGATGAGGTACCTGCAATGTCAGAACTGTCATGCACTCCAAAGTCCAGTTAAGGGAAATCCTGTTCCTGCCACCACCATCTTGTCTCACATATGATAGCAACTTTTGAGGCAGGACA
>MM9:10:18519282-18519327
ATACATCTCTTGAGAGCCCTAAGTAAGAGCAAGACAGACCTCCAGT
>MM9:10:18519331-18519521
CATGTAGGTTTCAGAAAAACAGTGAAAGCCCTTCAAGGATGCCAATACCCAGTACATCTTTCTTATGCACAGAAAGAACCTTACCTCATAGCTCACTCATAGCCGTCACCATTTGAACTTCAAGCAACTTCCAATTTGAACTCAATGGAAGCTTGCTGGT
GGTGCCAAAGAGTGCAATGCCACCTGACTTA
>MM9:10:18519523-18519641
TTCAGGTCTGCTGGCCACTTTGCTTAAGCAGAGTCTCTTACTGGCCTATAGTTTCTAAGTAGATTAGGCTAGCCATCCGATGAGTCCCAGGAATACACACCTGTTGGCTCACCTTGCAC
>MM9:10:18519649-18519693
GGTGTTCTTTCACTGTACTTCCCGGCCTCACACTTGCGATGGAGG
>MM9:10:18519725-18519969
TATGGTTTACTTCATTCGTCTTTCATTCACTAGGTTCCTGGTTCCTTTTCACATTTTGGTCTTTGTAGATTATCTTCATTCTAATTTTCATTTTTGTTTGAGACTTTTGTTCAAATTTTTAAAGGTTATATTTTGGTATTTCTTTTACTTTCTTTCAGTC
TCTTTTTTTCCTTGTAATATATTTCTTTGCCTCTAGTAATACTCTGGATTCTGTGGAATTTACTACAATTTGGCATTTAGTTAGG
```

Transcriptional units from RNA-seq



...



Annotating small RNA-seq libraries

- Identify expressed transcripts from trace read alignments to the target genome
- Determine what small RNA components are present
  - Screen for well known structural RNAs (e.g. ribosomal RNA, tRNAs, snoRNAs, etc)
  - Align transcripts to current version of miRbase to identify expressed microRNAs
  - Align transcripts to our own piRNA database built from recently published candidate piRNA sequences
- Set remaining unknown transcript population aside, examine for potentially novel RNAs

High-ranking miRbase alignments

KS35 (Spermatocytes)		KS45 (Round Spermatids)	
microRNA	Depth	microRNA	Depth
mmu-miR-805	1124	mmu-miR-184	5026
mmu-miR-191	472	mmu-miR-28	2799
mmu-miR-298	307	mmu-miR-423-5p	2083
mmu-miR-107	295	mmu-miR-470	494
mmu-miR-99b	290	mmu-miR-191	462
mmu-miR-28	255	mmu-miR-10b	411
mmu-miR-470	247	mmu-miR-34c	342
mmu-miR-151-3p	245	mmu-miR-182	320
mmu-miR-423-5p	220	mmu-miR-16	302
mmu-miR-881	220	mmu-miR-881	272
mmu-miR-184	195	mmu-miR-195	256
mmu-let-7d	108	mmu-miR-465c-5p	255
mmu-miR-34c	107	mmu-miR-743b-3p	161
mmu-miR-103	103	mmu-miR-151-3p	153
mmu-miR-743b-3p	87	mmu-miR-298	132
mmu-miR-202-5p	82	mmu-miR-107	130
mmu-miR-1196	81	mmu-miR-1195	130

KS35 (Spermatocytes)		KS45 (Round Spermatids)	
piRNA	Depth	piRNA	Depth
17446352.13	1364	17446352.13	1581
17446352.12	1322	17446352.12	1419
17446352.11	1201	17446352.11	1313
17446352.1	545	17446352.1	618
17446352.14	249	17446352.14	311
17446352.78	122	17446352.64	136
17446352.93	110	17446352.87	125
17446352.97	109	17446352.97	122
17446352.91	108	17446352.94	120
17446352.67	108	17446352.38	120
17446352.86	107	17446352.96	119
17446352.8	105	17446352.7	118
17446352.54	105	17446352.8	117
17446352.99	105	17446352.86	117
17446352.81	104	17446352.78	117
17446352.82	103	17446352.66	116
17446352.72	99	17446352.89	115

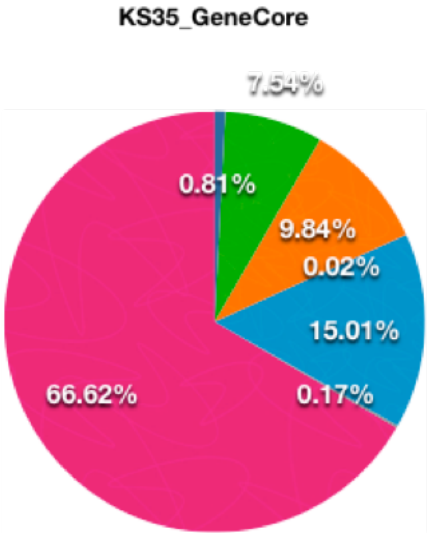
Composition of piRNA Database

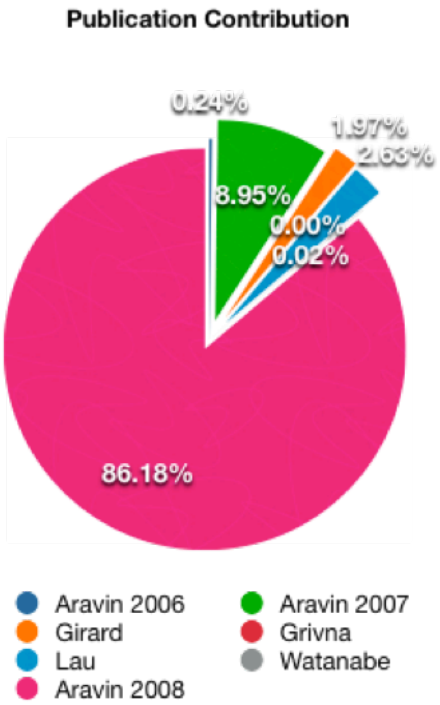
Total candidate sequences: 1,524,007

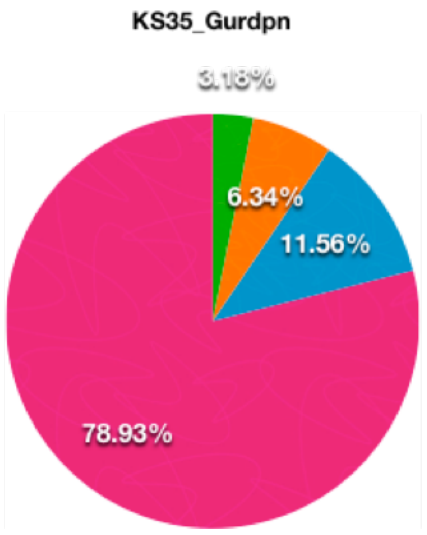
95%, 25nt	Number of piRNAs per publication (Tot=210,576)						
PubmedID	*16751777*	*17446352*	*16751776*	*16766680*	*16778019*	*16766679*	*18922463*
	Aravin 2006	Aravin 2007	Girard	Grivna	Lau	Watanabe	Aravin 2008
Total/dataset	3,638	136,417	30,024	40	40,102	355	1,313,431
KS45_GeneCore_eland	167	594	1,613	4	2,940	45	12,291
KS35_GeneCore_eland	150	1,390	1,815	3	2,768	31	12,286
KS35_Gurdon_eland	0	808	1,613	0	2,940	0	20,077
KS45_Gurdon_eland	159	1,283	1,907	5	2,888	33	23,371

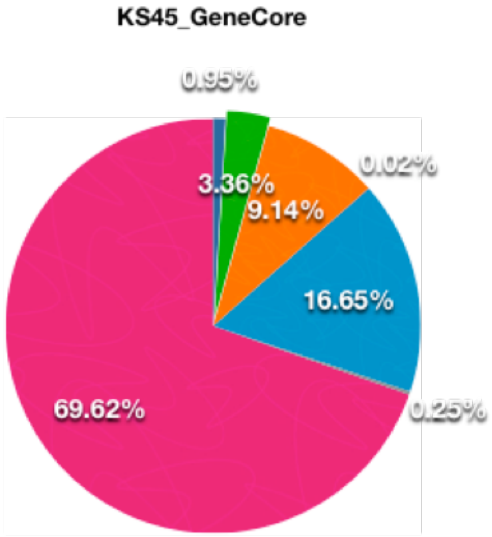
- Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, et al. (2006) Characterization of the piRNA complex from rat testes. Science 313: 363-367. PMID 16778019. Data came from Table S4. After sorting the table and taking only the uncharacterized sequences there were 40,102 piRNA candidates.
- Girard A, Sachidanandam R, Hannon GJ, Carmell MA (2006) A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature 442: 199-202. PMID 16751776. Data has been deposited into GenBank. There are 30,024 from this study.
- Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, et al. (2006) A novel class of small RNAs bind to MILI protein in mouse testes. Nature 442: 203-207. PMID 16751777. The piRNA candidate sequences are in an Excel table. It's supposed to be one of the files in the supplementary data, but is mislabeled on the website as S3 instead of S4. Removing the known sequences left 3,638 piRNA candidates from this study.
- Watanabe T, Takeda A, Tsukiyama T, Mise K, Okuno T, et al. (2006) Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. Genes Dev 20: 1732-1743. PMID 16766679. Data came from Table S7 (pdf). There are 355 candidate sequences.
- Grivna ST, Beyret E, Wang Z, Lin H (2006) A novel class of small RNAs in mouse spermatogenic cells. Genes Dev 20: 1709-1714. PMID 16766680. Data came from Table S1 (pdf). The sequences were only 40 of them.
- Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, Bestor T, Hannon GJ (2008) A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell 31: 785-799. PMID: 18922463. Data came from GEO, accession number GSE12757. There are 1,313,431 associated sequences.
- Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ (2007) Developmentally regulated piRNA clusters implicate MILI in transposon control. Science 316: 744-747. PMID 17446352. Data came from GEO, accession number GSE7414. There are 136,417 associated sequences.

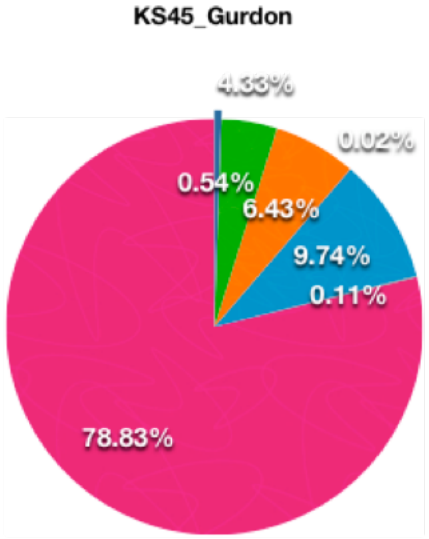
Composition of piRNA Database







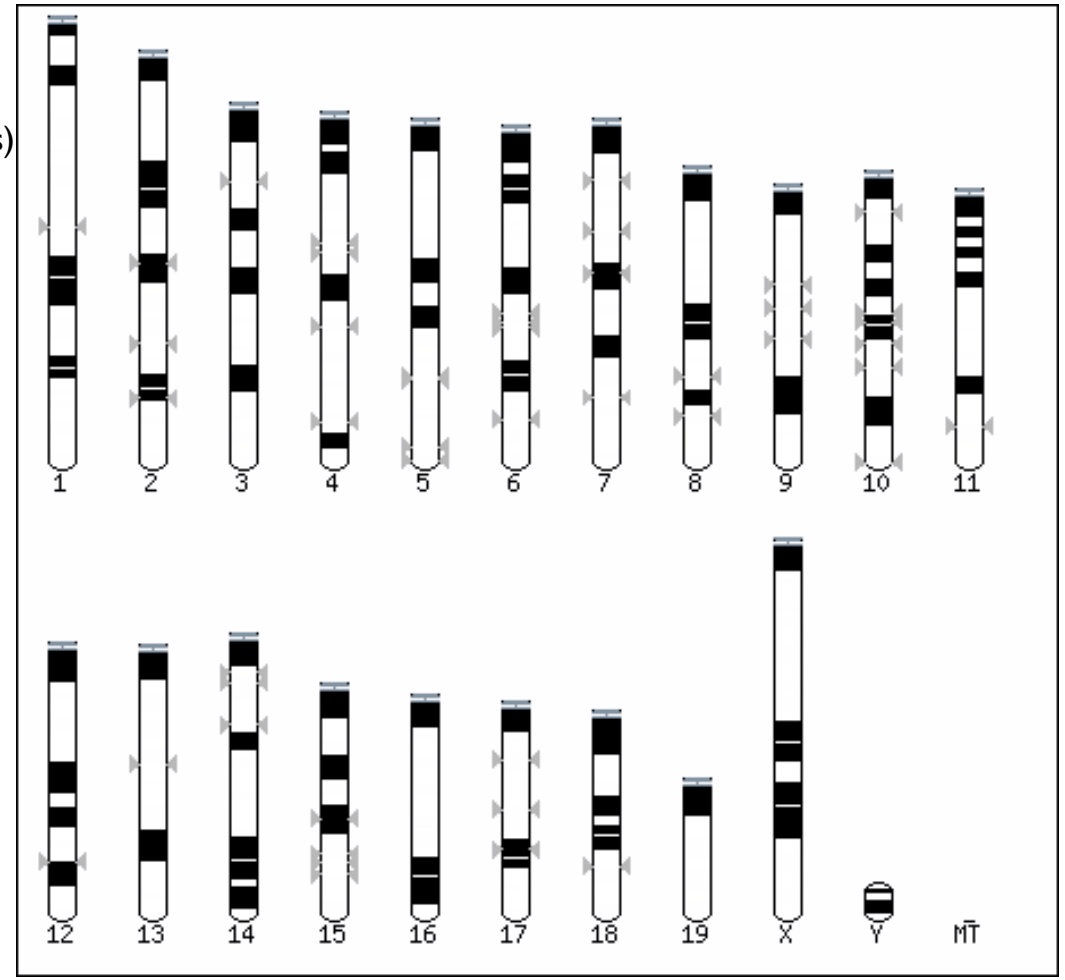




Karyogram (1)

Currently annotated piRNA clusters in mouse genome

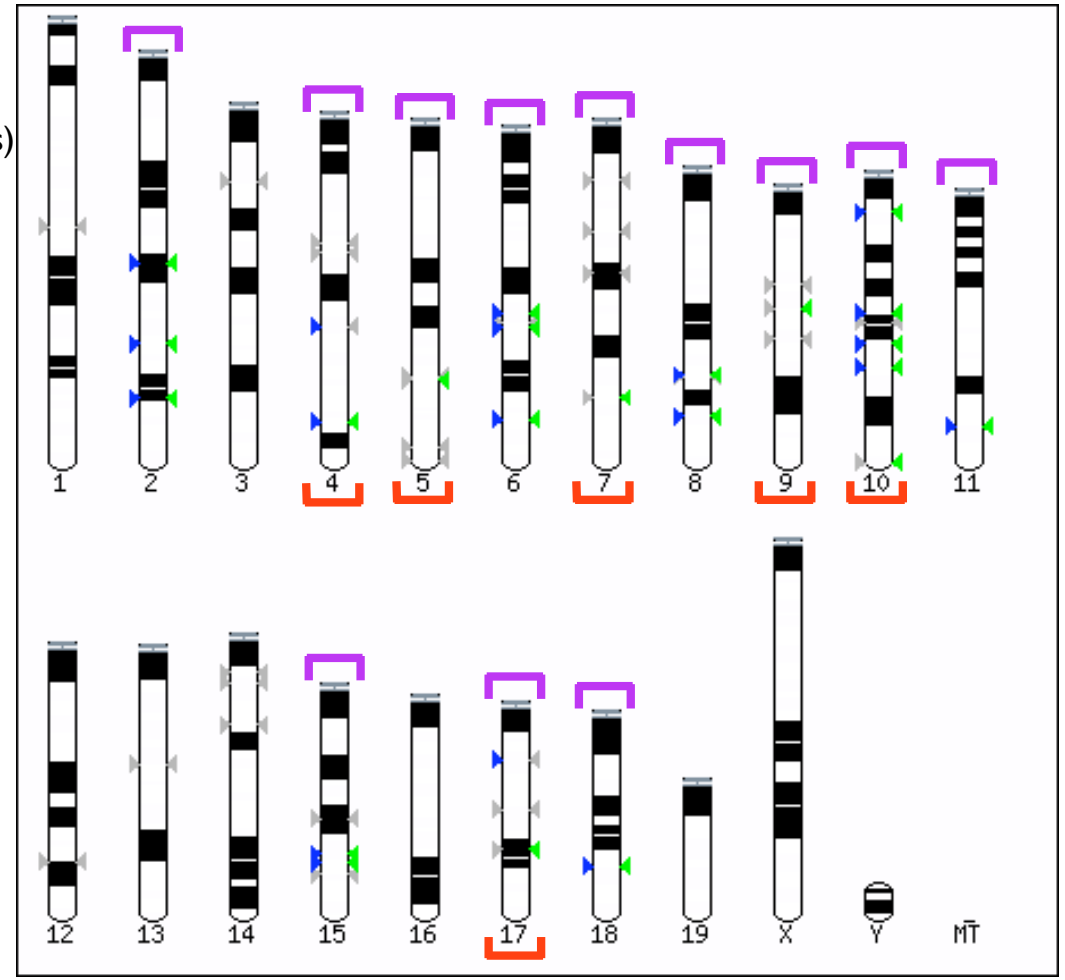
Left of chromosomes:
KS35 (Spermatocytes)
Right of chromosomes:
KS45 (Round Spermatids)



Karyogram (2)

Expressed transcripts within piRNA clusters (all levels)

Left of chromosomes:
KS35 (Spermatocytes)
Right of chromosomes:
KS45 (Round Spermatids)



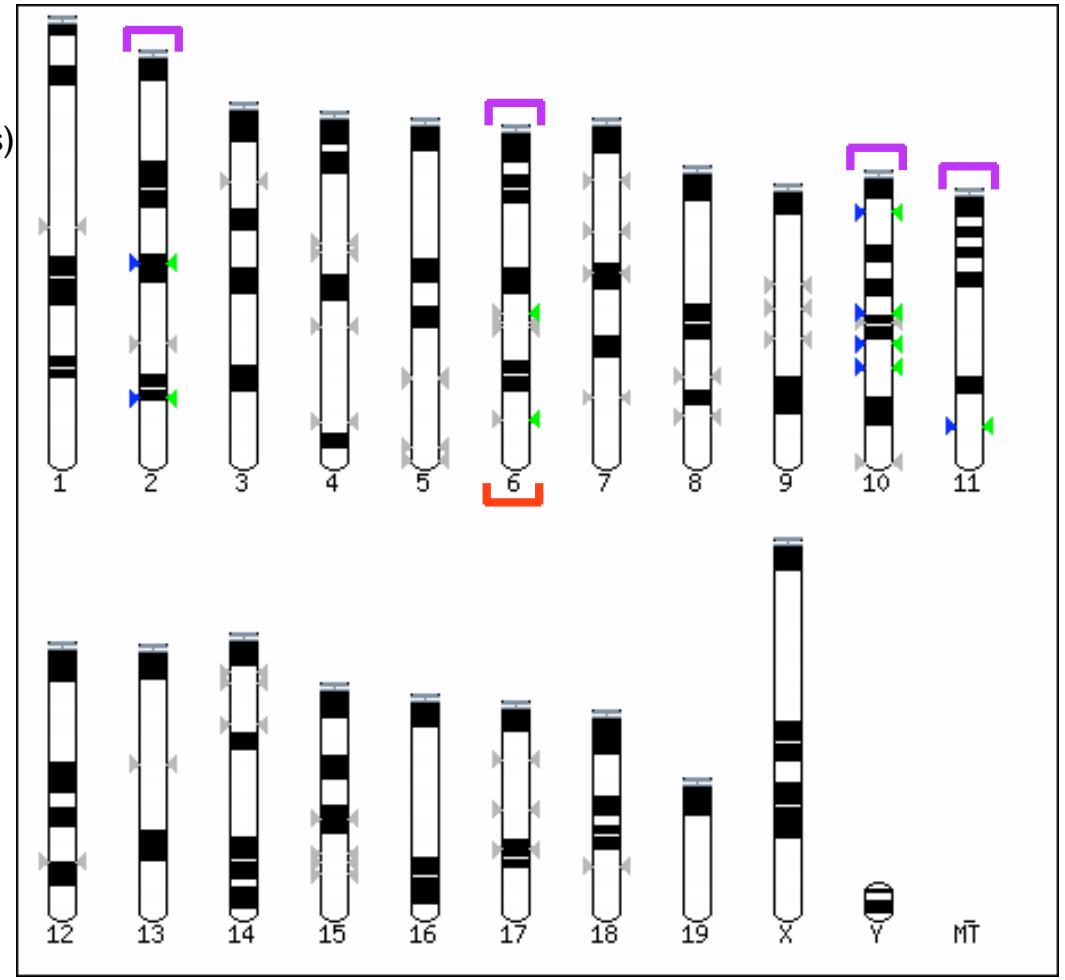
Expression observed
Differential expression observed

Karyogram (3)

Transcript abundance at mid-range levels

Left of chromosomes:
KS35 (Spermatocytes)

Right of chromosomes:
KS45 (Round Spermatids)



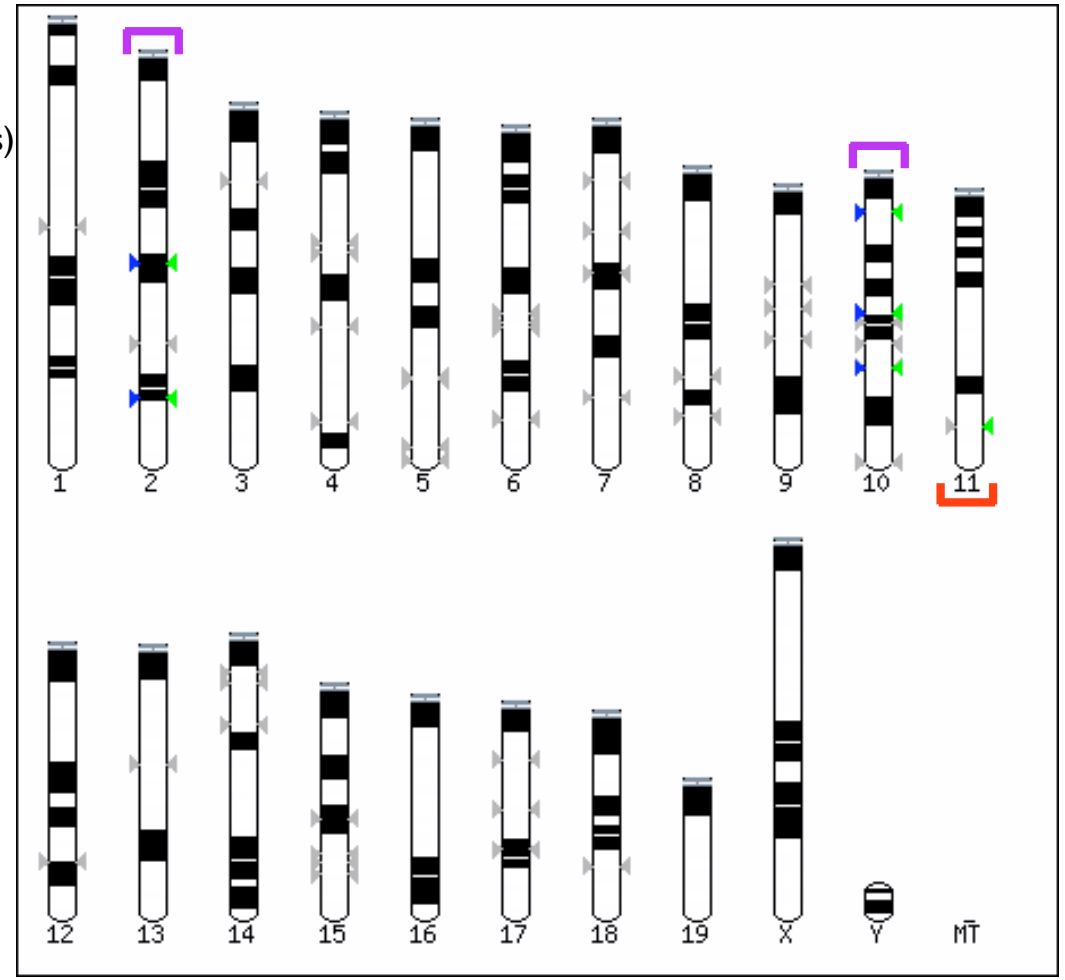
Expression observed

Differential expression observed

Karyogram (3)

Transcript abundance at high levels

Left of chromosomes:
KS35 (Spermatocytes)
Right of chromosomes:
KS45 (Round Spermatids)

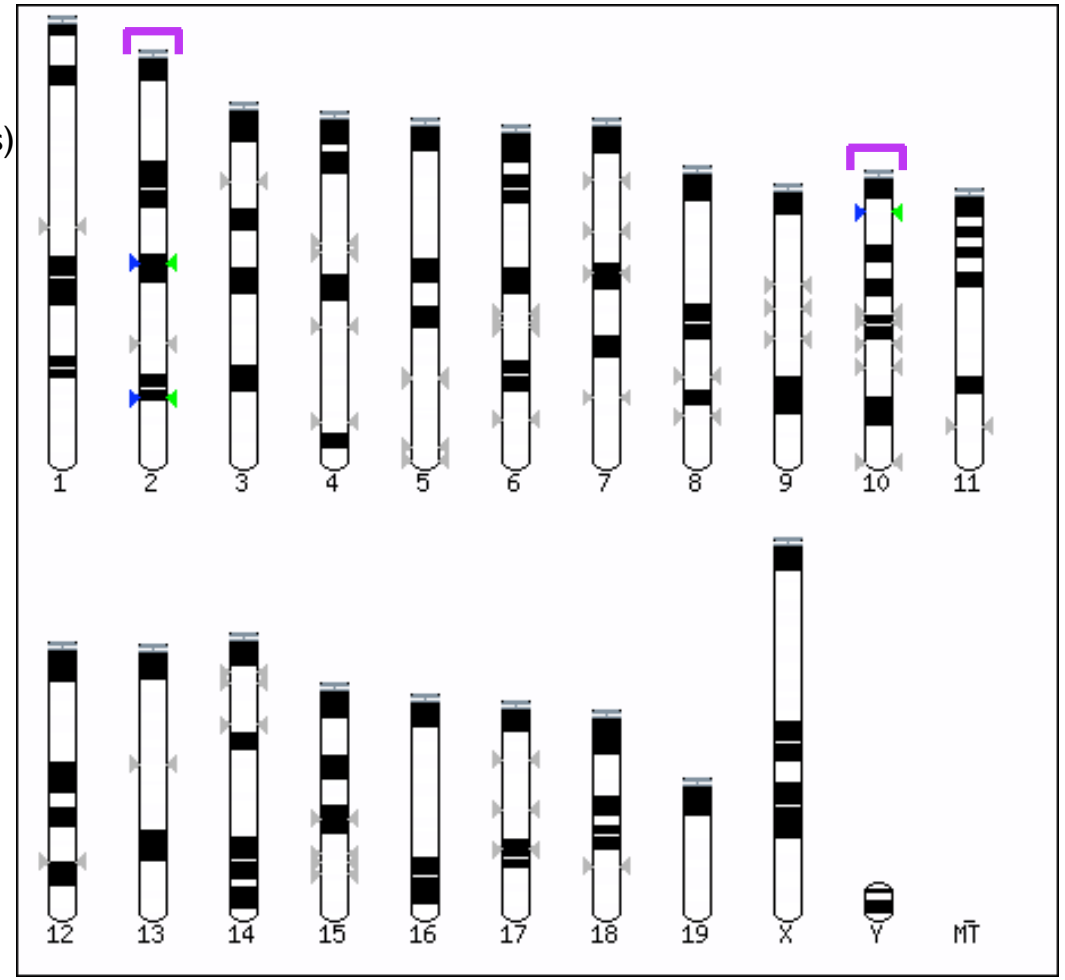


Expression observed
Differential expression observed

Karyogram (4)

Transcript abundance at very high levels

Left of chromosomes:
KS35 (Spermatocytes)
Right of chromosomes:
KS45 (Round Spermatids)



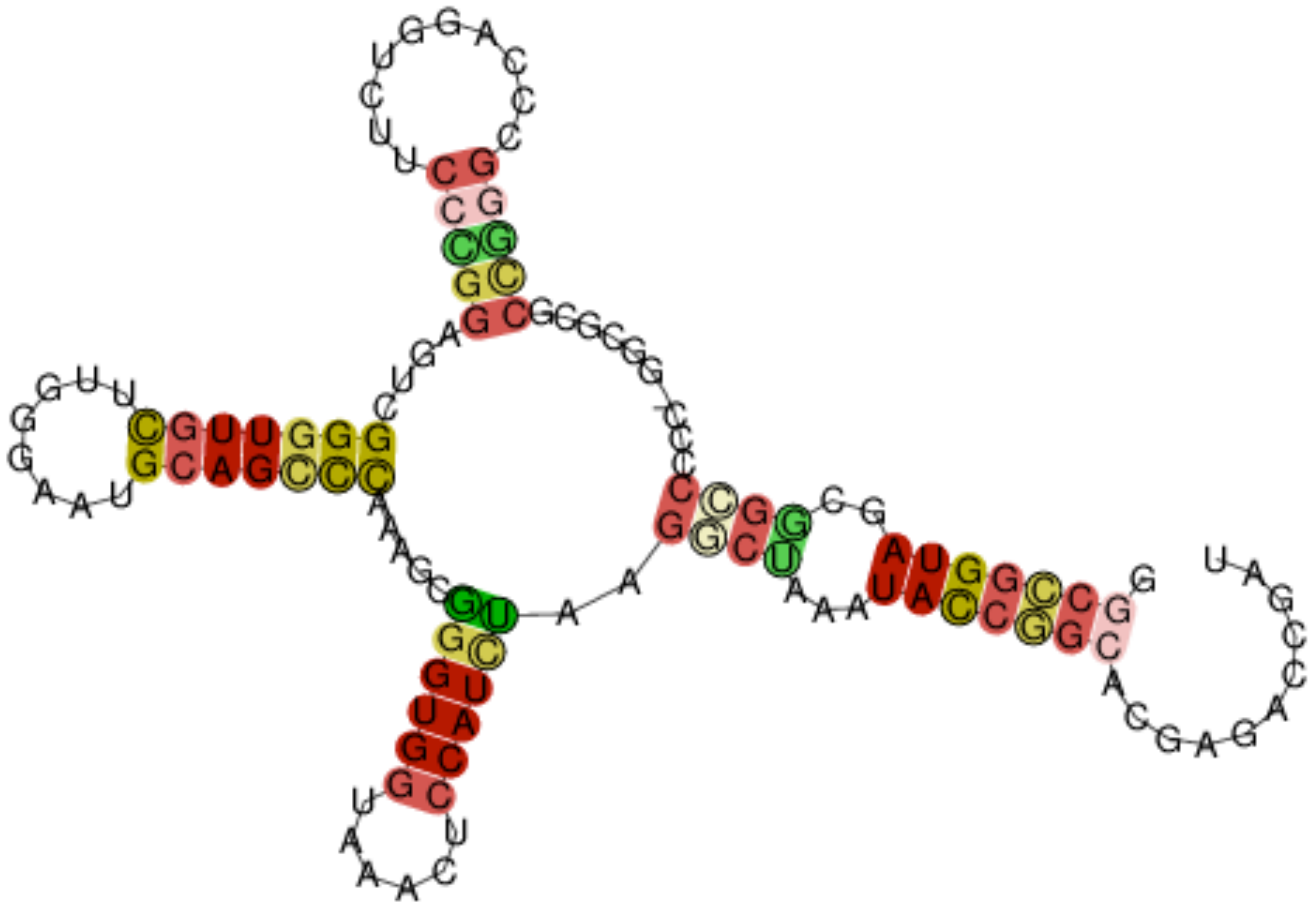
Expression observed
Differential expression observed

Analysis of novel RNA transcripts

- Transcribed regions fall into several categories
  - Correlate well with annotated (coding) gene loci
  - Correlate with existing non-coding RNAs
  - Novel transcripts

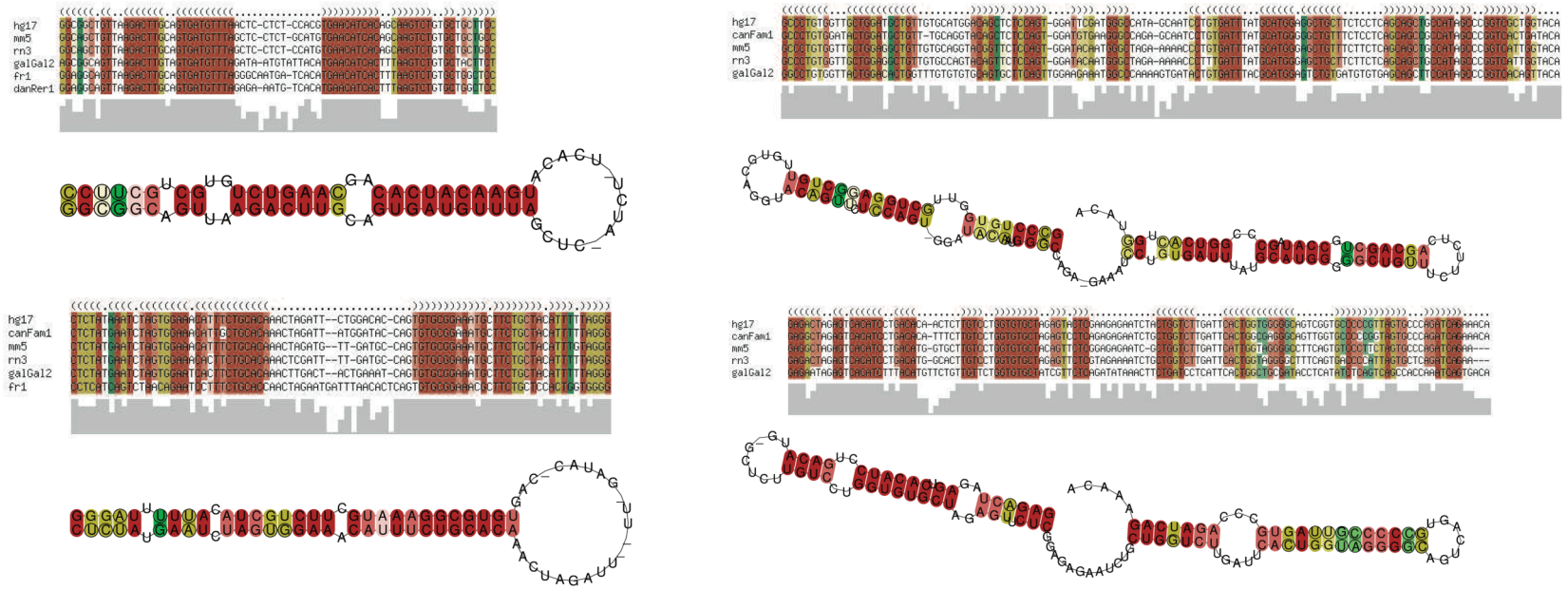
- Novel RNAs
  - To further characterize these, we perform RNA secondary structure prediction on thousands of candidate sequences
  - Look for favorable energy conformations
    - RNAfold (Vienna package), Mfold (Zucker lab)
  - Visualization of putative secondary structures
    - RNAplot (Vienna), StructureLab (Shapiro lab)
  - Homology across multiple species

Prediction of RNA Secondary Structure



Prediction of RNA Secondary Structure

Novel microRNA candidates conserved across species

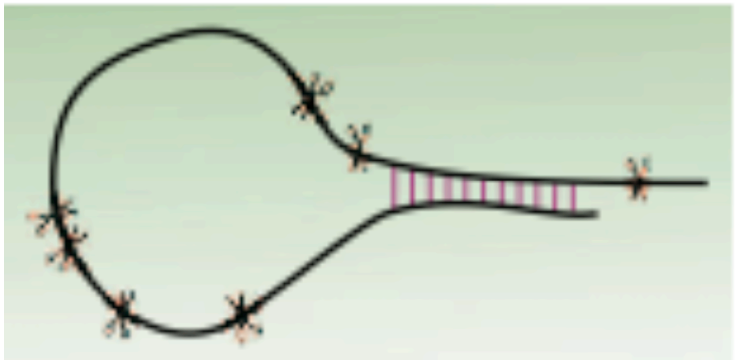


Stable hairpin consensus structures
Stem sequences are highly conserved
Loop sequences are divergent (variable)

Structural features of piRNAs

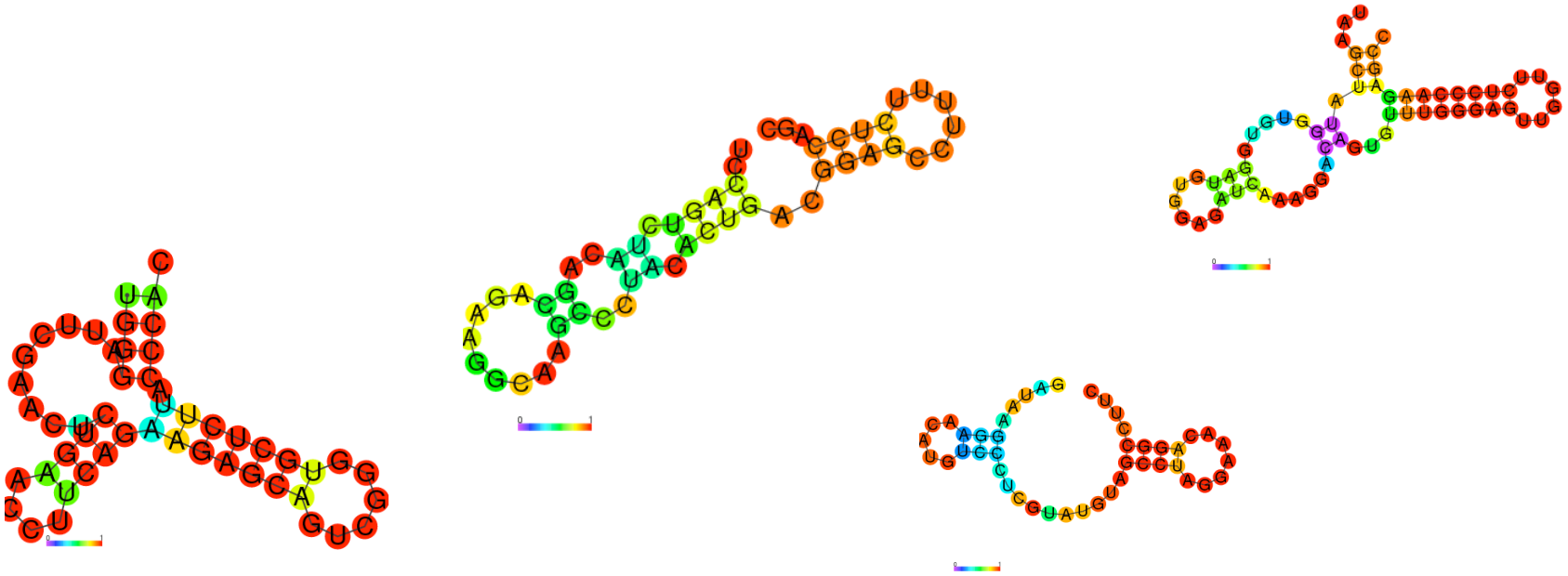
- As piRNAs are such a new class of regulatory non-coding RNA, their secondary structural properties are unknown

- Precursor transcripts are processed by a quasi-random mechanism
  - Weak sequence preference near the 5' U



Structural features of piRNAs

- Some structures can be identified based on features typically associated with microRNA hairpins
- It remains to be seen whether these will be characteristic of piRNAs as well



Summary

- Wide variety of RNA sequencing applications
- Library construction protocols differ according to the source material and aims of the experiment
- Open questions about strand specificity, level of coverage required for comprehensive transcriptome analysis
- Single- versus paired-end RNA sequencing
  - As read length increases, sequencing more single-end reads may be more informative